CHEST

Predicting Usual Interstitial Pneumonia Histopathology From Chest CT Imaging With Deep Learning

Alex Bratt, MD; James M. Williams, MD; Grace Liu, MD; Ananya Panda, MD; Parth P. Patel, BS; Lara Walkoff, MD; Anne-Marie G. Sykes, MD; Yasmeen K. Tandon, MD; Christopher J. Francois, MD; Daniel J. Blezek, PhD; Nicholas B. Larson, PhD; Bradley J. Erickson, MD, PhD; Eunhee S. Yi, MD; Teng Moua, MD; and Chi Wan Koo, MD

BACKGROUND: Idiopathic pulmonary fibrosis (IPF) is a progressive, often fatal form of interstitial lung disease (ILD) characterized by the absence of a known cause and usual interstitial pneumonitis (UIP) pattern on chest CT imaging and/or histopathology. Distinguishing UIP/IPF from other ILD subtypes is essential given different treatments and prognosis. Lung biopsy is necessary when noninvasive data are insufficient to render a confident diagnosis.

RESEARCH QUESTION: Can we improve noninvasive diagnosis of UIP by predicting ILD histopathology from CT scans by using deep learning?

STUDY DESIGN AND METHODS: This study retrospectively identified a cohort of 1,239 patients in a multicenter database with pathologically proven ILD who had chest CT imaging. Each case was assigned a label based on histopathologic diagnosis (UIP or non-UIP). A custom deep learning model was trained to predict class labels from CT images (training set, n = 894) and was evaluated on a 198-patient test set. Separately, two subspecialty-trained radiologists manually labeled each CT scan in the test set according to the 2018 American Thoracic Society IPF guidelines. The performance of the model in predicting histopathologic class was compared against radiologists' performance by using area under the receiver-operating characteristic curve as the primary metric. Deep learning model reproducibility was compared against intra-rater and inter-rater radiologist reproducibility.

RESULTS: For the entire cohort, mean patient age was 62 ± 12 years, and 605 patients were female (49%). Deep learning performance was superior to visual analysis in predicting histopathologic diagnosis (area under the receiver-operating characteristic curve, 0.87 vs 0.80, respectively; $P < .05$). Deep learning model reproducibility was significantly greater than radiologist inter-rater and intra-rater reproducibility (95% CI for difference in Krippendorff's alpha did not include zero).

INTERPRETATION: Deep learning may be superior to visual assessment in predicting UIP/IPF histopathology from CT imaging and may serve as an alternative to invasive lung biopsy.

CHEST 2022; 162(4):815-823

KEY WORDS: chest CT imaging; deep learning; idiopathic pulmonary fibrosis; interstitial lung disease; lung biopsy

FOR EDITORIAL COMMENT, SEE PAGE 734

ABBREVIATIONS: ATC = American Thoracic Society; AUC = area under the receiver-operating characteristic curve; CHP = chronic hypersensitivity pneumonitis; ILD = interstitial lung disease; IPF = idiopathic pulmonary fibrosis; NSIP = nonspecific interstitial pneumonitis; UIP = usual interstitial pneumonitis

AFFILIATIONS: From the Mayo Clinic, Rochester, MN.

CORRESPONDENCE TO: Alex Bratt, MD; email: bratt.alexander@mayo.edu

DOI: https://doi.org/10.1016/j.chest.2022.03.044

Take-home Points

Study Question: Can we improve noninvasive diagnosis of UIP by predicting ILD histopathology from CT scans using deep learning?

Results: Among 198 patients with biopsy-proven ILD, deep learning was superior to visual assessment in predicting UIP histopathology from CT scans and showed greater reproducibility.

Interpretation: Deep learning may improve noninvasive prediction of histopathologic UIP and could serve as an alternative to invasive lung biopsy.

Interstitial lung disease (ILD) is a heterogeneous group of diffuse lung diseases with variable treatment and prognosis depending on subtype.[1] Idiopathic pulmonary fibrosis (IPF) is a progressive, often fatal ILD characterized by the absence of a known or secondary cause and usual interstitial pneumonitis (UIP) pattern on chest CT imaging and/or histopathology. Distinguishing UIP/IPF from other ILD subtypes is essential and is achieved via multidisciplinary collaboration, taking into account history, physical examination, laboratory studies, pulmonary function test results, and high-resolution chest CT imaging.[2-4] In cases in which clinical and imaging features are insufficient to render a confident diagnosis, histopathologic evaluation via lung biopsy can provide additional diagnostic support, although at the cost of significant operative risk.[5,6] Our goal in the current study was to improve noninvasive diagnosis of IPF by predicting histopathologic UIP pattern from chest CT images.

In recent years, deep learning has shown promise as an aid to medical image interpretation. In some cases, deep learning models may perceive patterns that the human visual system cannot,[7] which opens up many possibilities to improve patient care. Classification of ILD from chest CT imaging is a particularly attractive target for deep learning given the often subtle or imperceptible differences in imaging patterns among ILD subtypes and relatively poor reproducibility among even highly trained experts.[8] Although a few prior studies have attempted to automate ILD pattern classification from CT images,[9-13] they are limited by smaller sample sizes, use of outmoded techniques, suboptimal diagnostic performance, and/or lack of histopathologic reference standards. We therefore sought to develop a deep learning model to improve classification of UIP vs non-UIP in cases undergoing lung biopsy. We hypothesized that deep learning may better assist in the recognition of atypically presenting radiologic patterns that correlate with underlying UIP histopathology, and therefore potentially defer or reduce the need to obtain a biopsy as part of the multidisciplinary diagnosis and workup.

Study Design and Methods

A waiver of approval was obtained from our Institutional Review Board for this retrospective study (Institutional Review Board #19-008965).

Patients and Data

Patients who had biopsy-proven ILD with chest CT scans between 1997 and 2020 were identified from an institutional registry and/or electronic medical record query (N = 1,239). We searched for patients exhibiting one of three histopathologic classes, namely UIP (n = 550), nonspecific interstitial pneumonitis (NSIP; n = 335), or chronic hypersensitivity pneumonitis (CHP; n = 354). All patients had been clinically evaluated at one of three medical centers (Mayo Clinic Rochester, Mayo Clinic Arizona, or Mayo Clinic Florida). The total data set size was 2,846 series (multiple series per CT scan) with 348,255 two-dimensional images.

Our data set contained one chest CT scan temporally closest to the biopsy date per patient. Contrast-enhanced and noncontrast CT scans from our own and outside institutions were used. Scanner parameters were variable, including 80 scanner models from seven manufacturers (Canon, General Electric, Hitachi, Imatron, Philips, Siemens, and Toshiba) and a mixture of high and low spatial frequency kernels. Mean slice thickness was 2.6 mm (range, 0.5-20 mm). Maximum-intensity projection images were included in the data set and accounted for the larger end of the slice thickness range. Image matrix size was 512 × 512. All CT scans were acquired in Digital Imaging and Communication in Medicine format and converted to Neuroimaging Informatics Technology Initiative format[14] for analysis.

CT scans were labeled as either UIP or non-UIP (CHP and NSIP) based on review of histopathology reports. Because ground truth labeling was solely based on histopathologic classification, not all cases of UIP were IPF by multidisciplinary diagnosis. The CT imaging data set was split into derivation (n = 1,038) and test (n = 201) sets, with the derivation set further subdivided into training (n = 894) and validation (n = 144). Each test set patient was manually reviewed. During this process, three cases were excluded, two for ambiguous histopathologic description and the other because the underlying lung disease had resolved by the time of the only available CT scan (Fig 1), yielding a total test set size of 198. Only axial CT images were included. Expiratory CT images were excluded to avoid class imbalance[15] as these were relatively uncommon. Prone series were not excluded because data augmentation included random vertical flip.

Deep Learning Model

We developed a custom deep learning pipeline to predict histopathologic diagnosis (UIP vs non-UIP) from chest CT imaging patterns (Fig 2), drawing inspiration from prior work.[10] First, the inferior 10% of axial slices (ie, the abdomen) was removed, and the remaining slices were split into 10 equally sized slabs. A single two-dimensional image slice was then randomly selected from each slab, generating a 10-slice montage. Two-dimensional images were then

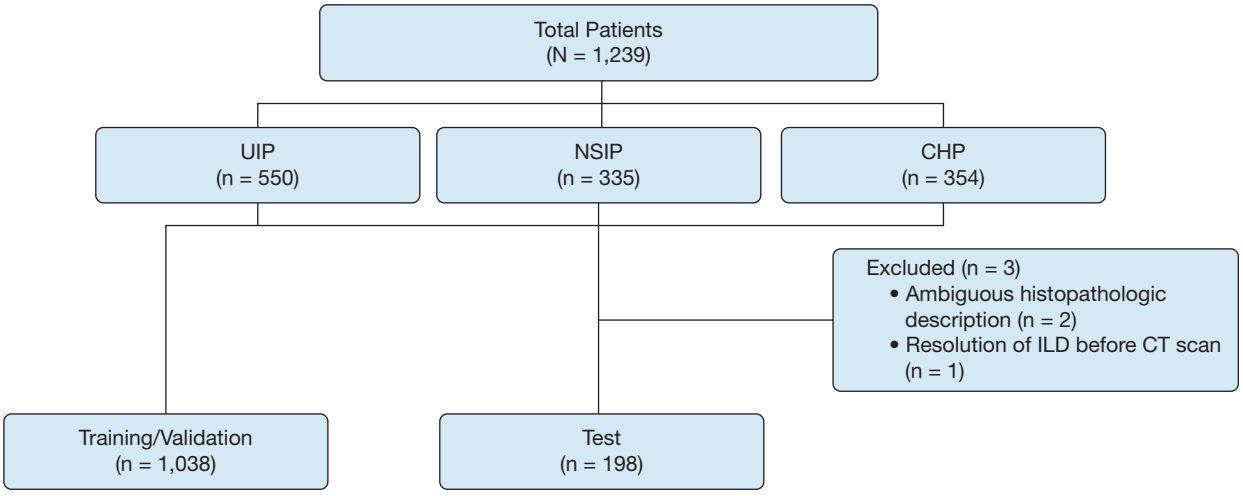

Figure 1 – *Data set flowchart. CHP = chronic hypersensitivity pneumonitis; ILD = interstitial lung disease; NSIP = nonspecific interstitial pneumonitis; UIP = usual interstitial pneumonitis.*

scaled to 400 × 400 and center cropped to 380 × 380. A lung window was applied (window 1500, level -600 Hounsfield units), and pixel intensities were rescaled between zero and one. Each preprocessed slice was then fed into an identical instance of EfficientNet-B3 (a commonly used two-dimensional deep learning model architecture),[16] creating a 64-element output vector. The output vectors for each slice were then concatenated and the resultant 640-element vector passed through a fully connected layer to produce a two-element logit vector (UIP, non-UIP). We summed the logit vectors from 10 randomly chosen montages and subsequently applied a softmax function along the class dimension. The outputs from each series of a given CT scan were summed prior to softmax being applied. A numerical threshold could then be chosen above which UIP would be deemed present. Test set inference was performed twice for reproducibility assessment because each montage contained randomly selected slices.

Training the model used the same preprocessing steps as noted earlier except with the addition of aggressive data augmentation (performed at runtime), including random crop (380 × 380), affine transformation,

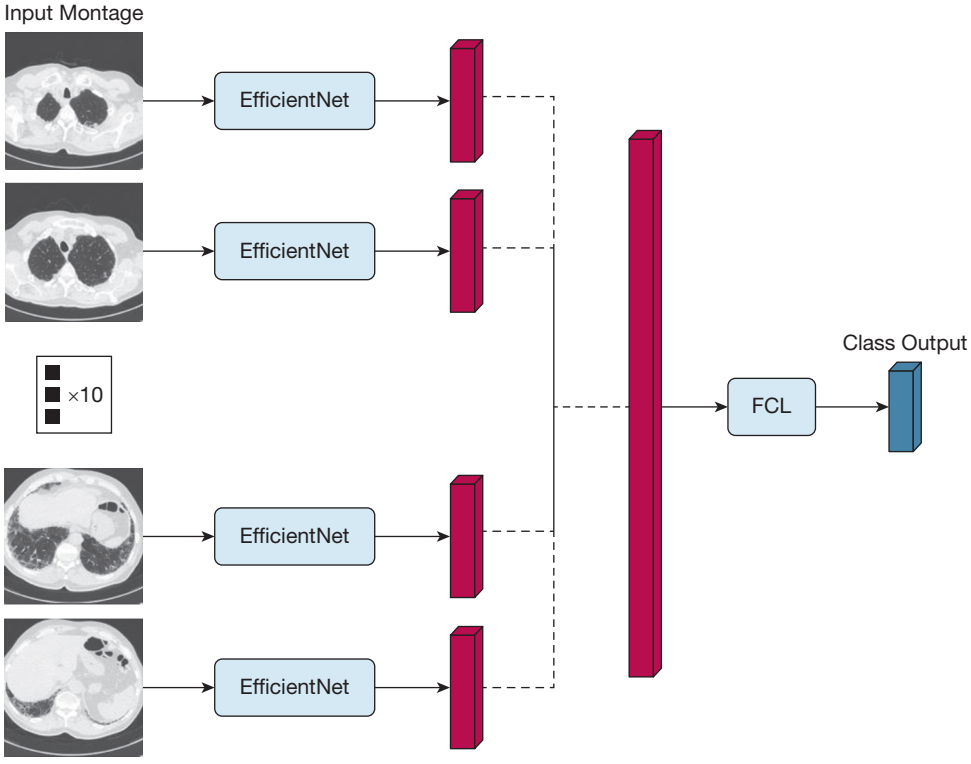

Figure 2 – *Deep learning model diagram depicting a 10-slice input montage and subsequent feedforward pathway. Dotted lines represent concatenation. FCL = fully connected layer.*

and addition of Gaussian noise. Incremental parameter updates were applied with the Adam method,[17] with the objective to minimize the cross entropy between output and true class. Batch size was six montages per training step (ie, 60 images) parallelized between two 24-gigabyte graphics cards. Performance of the model was monitored every 2,000 montages by calculating the area under the receiver-operating characteristic curve (AUC) over the entire validation subset. When it became obvious that performance had peaked, the model with the highest validation AUC was saved for testing. All model development and execution were performed done in PyTorch. Model code can be accessed at https://github.com/akbratt/UIP-DL.

Radiologist Classification

The test set was randomly split into two non-overlapping subsets of 99 cases. Each case was visually scored by two of four fellowship-trained thoracic radiologists (3-24 years' post-fellowship; L. W., A.-M. G. S., Y. K. T., and C. J. F.) on a four-point ordinal scale corresponding to the 2018 American Thoracic Society (ATS) guidelines[2] in which 1 = UIP, 2 = probable UIP, 3 = indeterminate for UIP, and 4 = alternative diagnosis. Raters had access to all series and images for one scan per patient but were blinded to other imaging and clinical history (other than age and sex visible in our picture archiving and communication system). One radiologist (L. W.) repeated ratings after 4 weeks for intra-rater reproducibility. All image review was performed on clinical workstations.

Deep learning model performance was evaluated against the mean of both radiologists' rating scores for each patient as well as every permutation of individual ratings (two test subsets and two raters per subset yielded four unique permutations of ratings).

Statistical Analysis

Data set size was determined solely by the number of patients in our institutional database meeting inclusion criteria. Diagnostic performance was assessed by means of receiver-operating characteristic curve analysis, with AUC as the primary metric. The paired DeLong test was used to compare AUC values.[18] For all comparisons, a two-sided P value threshold of .05 was considered statistically significant. Agreement between raters was assessed by using Krippendorff's alpha to address the partially crossed design, which was computed under both nominal and ordinal scoring. Similar methods were used for intra-rater reliability and binary agreement (UIP vs non-UIP) under different ATS thresholds. To compare reproducibility between pairs of binary classification samples, we used a bootstrapping strategy based on the method of Vanbelle and Albert,[19] modified for Krippendorff's alpha (bootstrap sample, N = 2,000). The Python packages NumPy[20] and SciPy[21] were used for statistical calculation.

Results

Table 1 presents the demographic features. Briefly, in the entire cohort, mean patient age at the time of CT imaging was 62 years, and 605 patients were female (49%). In the test set, mean age was 61 years, and 92 patients were female (46%). Approximately 12% of studies were contrast enhanced (n = 162). Of 1,236 patients meeting inclusion criteria, 830 (approximately 67%) were seen at Mayo Clinic Rochester, 277 (approximately 22%) at Mayo Clinic Arizona, and 129 (approximately 10%) at Mayo Clinic Florida. After preprocessing, deep learning model execution took 1.0 second per case on average. Figure 3 presents a histogram of mean ATS scores in the test set (mean score, 3.2).

Classification Performance

Radiologist and deep learning model performance are depicted in Figure 4. AUC for predicting histopathologic UIP was 0.87 (95% CI, 0.82-0.92) for deep learning vs 0.80 (95% CI, 0.75-0.86) for mean ATS class ($P = .03$). AUC values for different permutations of individual radiologist ratings were between 0.70 and 0.78 (all, $P < .05$ with respect to deep learning AUC). Individual radiologist sensitivity and specificity values at different classification thresholds are shown in Table 2. Radiologist classification performance in the current study was similar to slightly worse than previously reported, with mean specificity and sensitivity values of 0.97 and 0.22 for ATS score category < 2, respectively, compared with values of 0.97 and 0.38 described in previous work.[8] Mean ATS class is used for all further analysis because it showed greater AUC than any individual radiologist. This strategy gives human performance the greatest possible advantage against deep learning.

Figure 5 presents an example CT image that was correctly classified by the deep learning model as UIP (model output UIP softmax score, 1.00000) but classified by two radiologists as "alternative diagnosis." There were no cases that were correctly classified as typical UIP by

TABLE 1] Patient Characteristics

Characteristic	Full Cohort (N = 1,236)	Training Set (n = 894)	Validation Set (n = 144)	Test Set (n = 198)
Age, mean ± SD, y	62 ± 12	62 ± 12	61 ± 11	61 ± 11
Female, No. (%)	605 (49)	448 (50)	65 (45)	92 (46)
UIP, No. (%)	550 (44)	372 (42)	72 (50)	106 (54)
ΔT ± SD, d	188 ± 549	199 ± 591	138 ± 379	173 ± 447

ΔT = absolute time interval between CT scan and biopsy; UIP = usual interstitial pneumonitis.

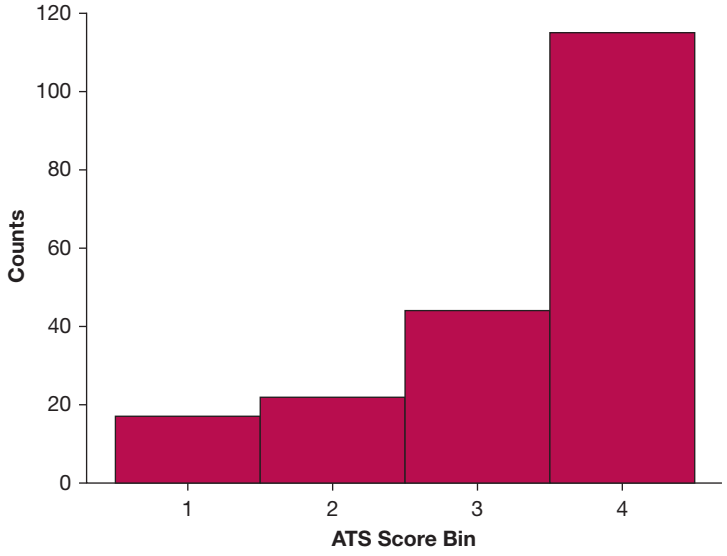

Figure 3 – *Histogram of mean ATS scores in the test set (1 = usual interstitial pneumonia [UIP]; 2 = probable UIP; 3 = indeterminate for UIP; 4 = alternative diagnosis). ATS = American Thoracic Society.*

both human raters but incorrectly classified as non-UIP by the deep learning model. One case, however, was correctly classified as alternative diagnosis by both raters but confidently and incorrectly classified as typical UIP by the deep learning model (output UIP softmax score, 1.00000; Figure 6). We do not provide heatmaps or use other forms of model output attribution such as Grad-Cam[22] because these are known to be unreliable.[23-26]

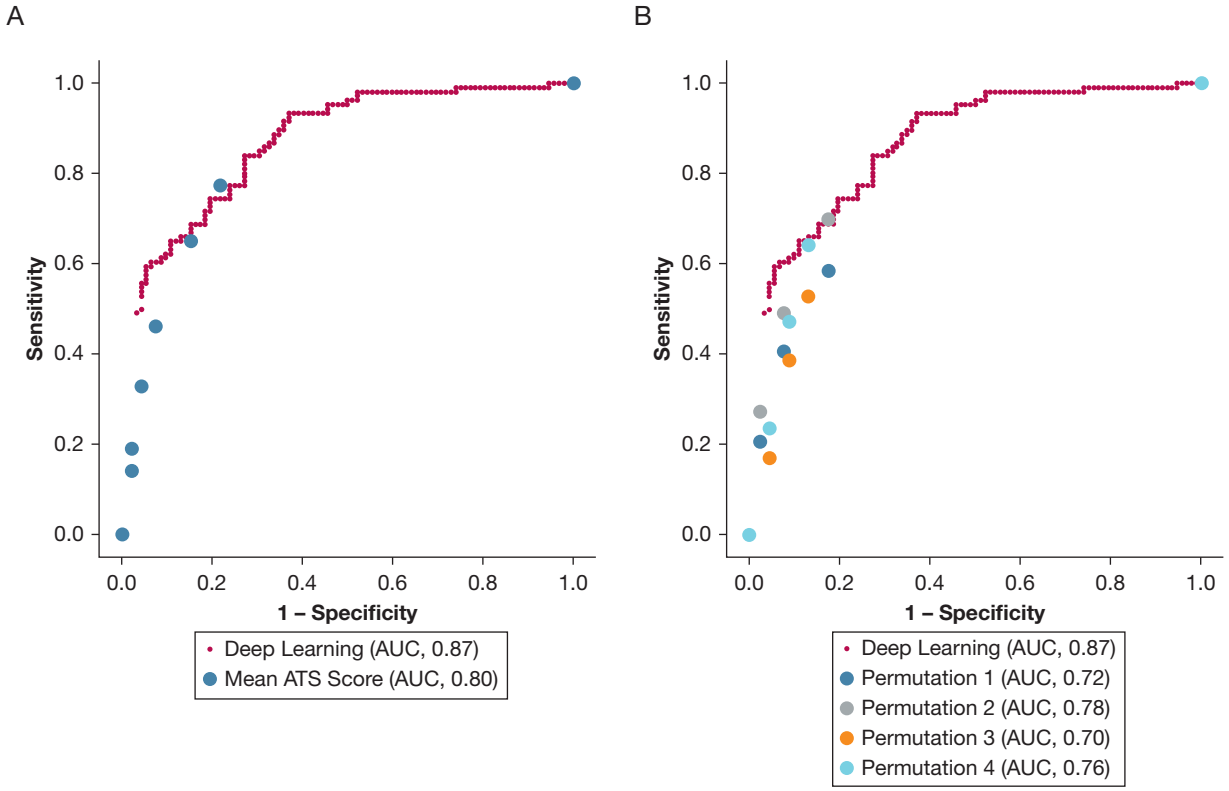

Figure 4 – *A-B, Diagnostic performance of the deep learning model vs manual evaluation. A, Deep learning performance vs mean ATS score. B, Deep learning performance vs different permutations of ATS score ratings. ATS = American Thoracic Society; AUC = area under the receiver-operating characteristic curve.*

TABLE 2] Radiologist Classification Performance at Three ATS ILD Score Thresholds

ATS Threshold	Performance Measure	Reader 1	Reader 2	Reader 3	Reader 4	Mean Performance[a]
< 2	Specificity	0.98 (0.87-1.0)	0.98 (0.87-1.0)	0.98 (0.87-1.0)	0.93 (0.81-0.98)	0.97 (0.92-0.99)
	Sensitivity	0.10 (0.04-0.22)	0.32 (0.20-0.46)	0.23 (0.13-0.37)	0.24 (0.14-0.38)	0.22 (0.10-0.38)
< 3	Specificity	0.94 (0.81-0.98)	0.91 (0.78-0.97)	0.94 (0.81-0.98)	0.89 (0.75-0.96)	0.92 (0.86-0.96)
	Sensitivity	0.37 (0.24-0.51)	0.44 (0.31-0.59)	0.54 (0.40-0.68)	0.41 (0.28-0.55)	0.44 (0.36-0.52)
< 4	Specificity	0.89 (0.76-0.96)	0.76 (0.60-0.87)	0.89 (0.76-0.96)	0.84 (0.70-0.93)	0.85 (0.71-0.94)
	Sensitivity	0.46 (0.33-0.60)	0.70 (0.56-0.82)	0.69 (0.55-0.81)	0.59 (0.45-0.72)	0.62 (0.48-0.74)

ATS = American Thoracic Society; ILD = interstitial lung disease.
[a]Mean performance denotes the mean of the performance metric, not the performance of the mean of two radiologist ATS scores.

Multi-Class Radiologist Reproducibility

Manual inter-rater reproducibility was moderate (alpha = 0.57) under the ordinal scaling of ATS class and modest under the nominal scale (0.40). Intra-rater reproducibility was high (0.73) under ordinal scaling and moderate under the nominal scale (0.59).

Binary Reproducibility

Radiologist reproducibility was also assessed by using binary classification (UIP vs non-UIP) at various ATS score thresholds. At a score threshold of 1 (1 = UIP; 2, 3, and 4 = non-UIP), inter-rater agreement was moderate (alpha = 0.59). At a score threshold of 2 (1 and 2 = UIP; 3 and 4 = non-UIP), agreement was also moderate (0.42). Intra-rater alpha was 0.71 at an ATS score threshold of 1 and 0.66 at a threshold of 2. Using softmax score thresholds of 0.9999996 and 0.9999451 (chosen to match the specificity values of the aforementioned ATS score thresholds), deep learning model reproducibility was near perfect (alpha = 0.95 and 0.98, respectively). The 95% CIs of the differences between binary ATS alpha and deep learning alpha values (at matched specificity values) did not include zero, indicating statistical significance (95% CIs of 0.189-0.554 and 0.436-0.742).

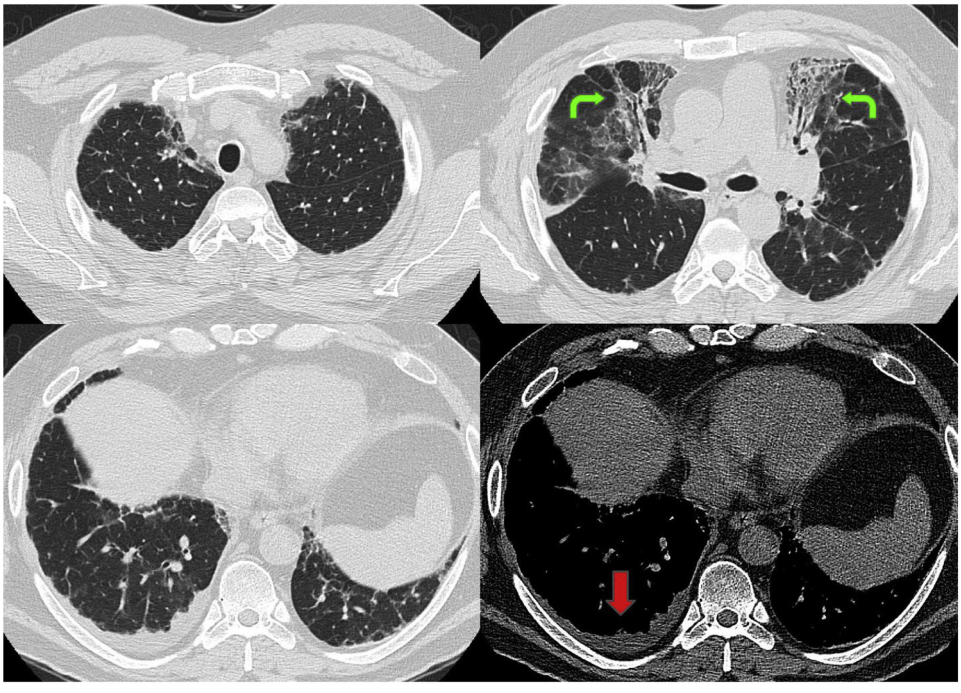

Figure 5 – *Selected slices from a CT scan correctly classified as usual interstitial pneumonitis by the deep learning model but incorrectly classified by two expert radiologists. This case is challenging because it demonstrates visually atypical features for usual interstitial pneumonitis/idiopathic pulmonary fibrosis, namely mid-lung and peribronchovascular-predominant fibrosis (green curved arrows) as well as pleural effusion (red straight arrow). Note that the lung apices and bases are relatively spared.*

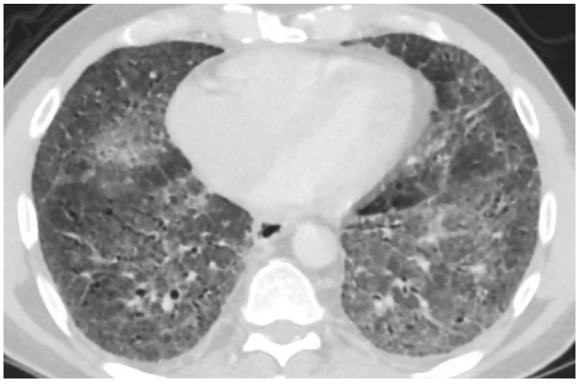

Figure 6 – *Representative slice from a CT scan correctly classified as compatible with a non-idiopathic pulmonary fibrosis diagnosis by both human raters but confidently labeled as typical usual interstitial pneumonitis pattern by the deep learning model.*

Discussion

To our knowledge, this study is the first to show superhuman performance in classifying UIP histopathology from chest CT scans using deep learning. In the setting of IPF, ILD classification is important to inform prognosis, guide treatment, and determine eligibility for clinical trials. Unfortunately, the true etiology of ILD often remains elusive even following extensive noninvasive testing.[27,28] In this setting, lung biopsy may improve diagnostic confidence but incurs a nonnegligible risk of operative morbidity[5,6] We sought to improve noninvasive assessment of atypically presenting IPF by predicting UIP histopathology from CT scans. In this retrospective study, we showed that our deep learning model provides higher diagnostic performance than visual evaluation (AUC, 0.87 vs 0.80; $P = .03$) with greater reproducibility. Our model may reduce the need for invasive lung biopsy in patients with atypical or less evident UIP CT imaging patterns by increasing confidence in underlying UIP histopathologic diagnosis.

The current study builds upon previous work using deep learning to evaluate ILD. Walsh et al[10] and Christe et al[11] developed deep learning models to classify ILD patterns on CT scans based on established guidelines,[2,3] reporting similar performance with respect to radiologists. We expand on this work by using histopathology as a reference standard rather than CT image pattern, enabling us to show that deep learning may actually outperform experienced radiologists to better support multidisciplinary diagnosis without need for biopsy. Shaish et al[9] reported modest deep learning performance in classifying histopathologic ILD subtype using "virtual wedge resection." The current study has four key advantages with respect to this work. First, we achieved greater diagnostic performance (AUC, 0.87 vs 0.74),[9] which we attribute primarily to the use of entire CT images rather than small peripheral subvolumes. Because ILD is a diffuse lung disease, it is possible that important information is lost when omitting the central lung parenchyma. Second, we provide direct comparison with human performance, showing that our model may be superior to an established visual chest CT imaging classification scheme. Third, our model requires many fewer preprocessing and postprocessing steps (eg, lung mask generation, false-positive reduction), which decreases model complexity and execution time, further improving robustness and applicability. Finally, relative to the single institution-based algorithm described in another study,[9] our model is likely more generalizable because it was trained on a heterogeneous data set of scans from multiple institutions with varied scanner manufacturers and models.

The current study has several limitations. First, the data set only includes three pathologically proven disease entities (UIP, NSIP, and CHP). However, these are the most common histologic patterns clinically and radiologically overlapping with IPF, requiring additional histopathology to confirm in many. Model performance is unknown for other ILD subtypes and may be addressed in future prospective studies. Second, the current cohort only includes patients who underwent lung biopsy, which is heavily biased toward diagnostically challenging cases. Because the majority of classic UIP cases can be confidently diagnosed with noninvasive testing and thus do not require biopsy, the incremental benefit from a deep learning solution is greatly diminished. Nevertheless, assessing our model in classic UIP cases is of relevance, particularly as the pattern may be found in clinically diverse diseases with subtle but unrecognized differences (IPF vs early or undiagnosed rheumatoid arthritis-related fibrotic ILD, for example, may have similar classic UIP CT imaging patterns). Third, we do not intend this model to be executed on data outside its training distribution, given the well-known limitations of deep learning model generalization.[29,30] Thus, we do not report model performance on external, unseen data (data from all three centers were included in the training set). Others wishing to evaluate the current model in their own practice are encouraged to train a randomly initialized instance of our model on local data. Fourth, this study has all the inherent limitations associated with

retrospective analysis, which we plan to address in future prospective studies. Fifth, deep learning model performance was only benchmarked against four radiologists, a number which may not fully represent the entire spectrum of human capability. Finally, this model does not replace the need for multidisciplinary diagnosis. Although we used histopathology as the reference standard, it is not the sole determinant of IPF diagnosis. Recognizing the importance of collaboration in diagnosing IPF, the next logical step will be to evaluate the utility of our model in a multidisciplinary setting.

Interpretation

We provide multicenter validation of a deep learning model that may improve CT imaging prediction of histopathologic UIP in patients with ILD. This can increase noninvasive diagnostic confidence and thereby provide an alternative to invasive lung biopsy.

Acknowledgments

Author contributions: A. B. is the guarantor of the study. The author contributions were as follows: study design, A. B. and C. W. K.; data acquisition, A. B., J. M. W., G. L., A. P., P. P. P., D. J. B., B. J. E., and C. W. K.; data annotation, L. W., A.-M. G. S., Y. K. T., and C. J. F.; deep learning, A. B.; data analysis, A. B., N. B. L., E. S. Y., T. M., and C. W. K.; and manuscript drafting, A. B. and C. W. K. All authors were responsible for manuscript revisions, final approval, and study accountability.

Funding/support: This study was funded by Mayo Clinic Internal Research Support.

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following: B. J. E. is a shareholder in VoiceIT Inc., FlowSIGMA Inc., and Yunu Inc. (unrelated to current project). None declared (A. B., J. M. W., G. L., A. P., P. P. P., L. W., A.-M. G. S., Y. K. T., C. J. F., D. J. B., N. B. L., E. S. Y., T. M., C. W. K.).

Role of sponsors: The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

References

1. Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med*. 2018;378(19):1811-1823.
2. Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med*. 2018;198(5):e44-e68.
3. Lynch DA, Sverzellati N, Travis WD, et al. Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. *Lancet Respir Med*. 2018;6(2):138-153.
4. Cavazza A, Rossi G, Carbonelli C, Spaggiari L, Paci M, Roggeri A. The role of histology in idiopathic pulmonary fibrosis: an update. *Respir Med*. 2010;104:S11-S22.
5. Hutchinson JP, McKeever TM, Fogarty AW, Navaratnam V, Hubbard RB. Surgical lung biopsy for the diagnosis of interstitial lung disease in England: 1997-2008. *Eur Respir J*. 2016;48(5):1453-1461.
6. Hutchinson JP, Fogarty AW, McKeever TM, Hubbard RB. In-hospital mortality after surgical lung biopsy for interstitial lung disease in the United States. 2000 to 2011. *Am J Respir Crit Care Med*. 2016;193(10):1161-1167.
7. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861-867.
8. Shih AR, Nitiwarangkul C, Little BP, et al. Practical application and validation of the 2018 ATS/ERS/JRS/ALAT and Fleischner Society guidelines for the diagnosis of idiopathic pulmonary fibrosis. *Respir Res*. 2021;22(1):124.
9. Shaish H, Ahmed FS, Lederer D, et al. Deep learning of computed tomography virtual wedge resection for prediction of histologic usual interstitial pneumonitis. *Annals ATS*. 2021;18(1):51-59.
10. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med*. 2018;6(11):837-845.
11. Christe A, Peters AA, Drakopoulos D, et al. Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images. *Invest Radiol*. 2019;54(10):627-632.
12. Depeursinge A, Chin AS, Leung AN, et al. Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution CT. *Invest Radiol*. 2015;50(4):261-267.
13. Chung JH, Adegunsoye A, Oldham JM, et al. Vessel-related structures predict UIP pathology in those with a non-IPF pattern on CT. *Eur Radiol*. 2021;31(10):7295-7302.
14. Cox R, Ashburner J, Breman H, et al. A (sort of) new image data format standard: NIfTI-1: WE 150. *Neuroimage*. 2004;22.
15. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*. 2018;106:249-259.
16. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, PMLR. 2019;97:6105-6114.
17. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, eds. *3rd International Conference on Learning Representations, ICLR 2015*. Conference Track Proceedings; 2015. May 7-9, 2015, San Diego, CA.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
19. Vanbelle S, Albert A. A bootstrap method for comparing correlated kappa coefficients. *J Statistic Computat Simulation*. 2008;78(11):1009-1015.
20. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585:357-362.
21. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17:261-272.
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*. 2017:618-626.
23. Eitel F, Ritter, K; Alzheimer's Disease Neuroimaging Initiative (ADNI). Testing the robustness of attribution methods for convolutional neural networks *in MRI-based Alzheimer's disease classification. Interpretability of Machine Intelligence in Medical Image Computing and* Multimodal Learning for Clinical Decision Support. Springer; 2019:3-11.
24. Crosby J, Chen S, Li F, MacMahon H, Giger M. *Network output visualization to uncover limitations of deep learning detection of pneumothorax. Medical Imaging 2020: Image Perception, Observer*

Performance, and Technology Assessment, Vol. 11316. International Society for Optics and Photonics; 2020:113160O.

25. Young K, Booth G, Simpson B, Dutton R, Shrapnel S. Deep neural network or dermatologist? *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer; 2019:48-55.
26. Arun N, Gaw N, Singh P, et al. Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv 200802766*. Posted online July 14, 2021. Accessed August 19, 2021. http://arxiv.org/abs/2008.02766
27. Sverzellati N, Wells AU, Tomassetti S, et al. Biopsy-proved idiopathic pulmonary fibrosis: spectrum of nondiagnostic thin-section CT diagnoses. *Radiology*. 2010;254(3):957-964.
28. Yagihashi K, Huckleberry J, Colby TV, et al. Radiologic–pathologic discordance in biopsy-proven usual interstitial pneumonia. *Eur Respir J*. 2016;47(4):1189-1197.
29. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.
30. Yan W, Huang L, Xia L, et al. MRI manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for MR images acquired with different scanners. *Radiol Artif Intell*. 2020;2(4):e190195.